DOCUMENT RESUME

ED 054 208 TM 000 783

AUTHOR	Green, Donald Ross
TITLE	Biased Tests.
INSTITUTION	CTB/McGraw Hill, Monterey, Calif.
PUB DATE	71
NOTE	10p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Educationally Disadvantaged, Minority Groups, *Negroes, Predictive Ability (Testing), Scores, *Standardized Tests, *Test Bias, Test Construction, *Testing, *Testing Problems, Test Interpretation, Test Reliability, Test Selection, Test Validity

ABSTRACT

This paper is concerned with the accusations made by such groups as the Association of Black Psychologists in their call for a moratorium on testing, that standardized tests are biased. A biased test measures one trait in one group of people but a different trait in a second group. Evidence about the amount of bias in tests is thin. Bias must be determined by research on each instrument. A commitment to such research is in order. If bias is found, reasonable courses of action include test revision, alteration in interpretation, and discontinuance of testing. (Author/AG)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEH REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 054208

BIASED TESTS

by
Donald Ross Green
CTB/McGraw-Hill
Del Monte Research Park
Monterey, California 93940

There is a long-standing notion that tests may be biased against blacks and others (e.g. Pintner 1923, p. 343; Eells, Davis, Havighurst, Herrick & Tyler 1951). In recent years the charge has been made with increasing frequency and vehemence (e.g. Brown & Russell 1964; Wasserman 1969; Danielian 1969; Brazziel 1969; Williams 1970a, 1970b). For the most part these charges have either been ignored or dismissed as ill-founded both by the publishers of such tests and by the people and institutions who use them (e.g. Clemans 1970; Sommer 1970; Stanley 1971; Wrightstone 1969). Since the charge is basically that large numbers of people, especially children, have been systematically cheated and misused, vehemence and even anger seem appropriate it if is true. A biased test produces scores that mean different things for different groups and therefore, if the bias is not recognized, teachers make erroneous inferences about their teaching strategies and about their students; school administrators make erroneous inferences about their programs and curricula; school boards and legislators make erroneous inferences about the merit of their policies and personnel; and, heaven help us, the researchers and theoreticians develop false models and explanations of schooling and human nature. Is anger, then, an over-reaction?

Perhaps the most explicit and detailed set of charges against standardized tests are to be found in the two articles by Williams (1970a, 1970b). He states that published intelligence and achievement tests are biased against black people; he gives his reasons, describes the evidence he is accumulating to support them, and expresses his anger. He may be wrong in indicting all ability and achievement tests, but at least some of his statements cannot be refuted with available evidence and his own research may yet provide strong support for his attack. In short, his charges deserve serious consideration and, in fact, they have received it.

Two sets of replies have appeared. There were three responses to Williams' "Danger: Testing and Dehumanizing Black Children," which appeared in *The Clinical Child Psychology Newsletter* (Milgram 1970; Wikoff 1970; Newland 1970), and a rejoinder by Williams (1970c). At the invitation of the editor of *The Counseling Psychologist*, five test publishers responded to Williams' article in that journal (Messick & Anderson 1970; Bennett 1970; Sommer 1970; Clemans 1970; Munday 1970). These answers, like Williams' statements, deserve careful attention. But there is much more to be said and to be learned. A striking feature of the whole discussion about test bias is the scarcity of broad, solid evidence about the nature and generality of bias in tests in spite of the fact that most of the issues are open to empirical investigation.

Partly because of this lack of evidence, it does not seem possible to fully agree with either Williams or any one of the replies to him. In this paper I shall first try to define bias more clearly, secondly to discuss some of the points raised by Williams along with some of the replies already published, and finally to indicate briefly the steps being taken at CTB/McGraw-Hill to deal with the matter and learn more about it.

TM 000 783

It is useful to distinguish between biased tests and biased people. These are two logically separable problems, even though Williams seems convinced that tests are biased because biased people made them that way. Were he correct it nevertheless would be much easier to improve the tests than to change the people.

A test may be called biased if it measures something different in different groups of people, i.e., different subsets of the population for which the test was designed. Suppose, for example, there was a test which measured only verbal facility in one group and only anxiety in a second group. This test would be biased. It would be biased against the first group if used as a measure of anxiety but not if it were treated as a test of verbal facility. In either case, to use the test as a measure of the same attributes with both groups would provide faulty and misleading information about the members of one of the two groups. For this definition of bias to be useful, the groups cannot be determined by their scores on the test in question; usually people refer to groups based on ethnic origin, social class, sex, or the like, but any common set of extraneous characteristics will do. To reiterate, a test is biased when the behavior it elicits is different and has a different meaning for different groups; whenever bias is not recognized and/or allowed for, the scores do not mean what they appear to mean and unfair treatment of many people is a likely consequence.

Another illustration comes from an example offered by Williams. He states that the typical reading comprehension test is biased because it functions differently with white students on the one hand, and with blacks on the other. For white students the test probably functions as it was meant to; that is, it measures the students' ability to interpret a passage with prior knowledge playing a negligible role (Marks & Noll 1967). For many of the black children on the other hand, prior knowledge may play a major role in determining their scores: thus the test is only partly measuring comprehension. To drive his point home, Williams offers a passage which could be used to measure reading comprehension among urban blacks, but which would obviously function badly among suburban whites. Clemans (1970) and Sommer (1970) take Williams to task on this matter, pointing out that few, if any, teachers would use such material and hence that his test would not reflect the teacher's educational objectives. They are undoubtedly correct, but Williams has made his point well. For many black children, scores on a test using Williams' material would reflect their reading comprehension skill, while for most white children the scores would largely reflect their unfamiliarity with the terminology. In short, it would be a plainly biased test. It follows that those tests now in common use may well have the same kind of drawback when used with children growing up in subcultures which differ from the mainstream to any substantial degree.

Note that the systematic misuse of a test so that unfair consequences ensue for everyone is not test bias but rather the consequence of ignorance or bias in people. Such misuses of tests are all too widespread, but it does not help to identify the test as the culprit. This particular point is widely acknowledged and has been discussed at length in the published replies to Williams. Therefore the point need not be considered further here. Note also that the probability that a test is biased may or may not be readily determined by perusal; test bias can be demonstrated empirically only by conducting validity studies designed to examine that possibility.

The definition of test bias offered here is in line with that used by Messick and Anderson (1970) in their reply to Williams, and with definitions previously used by others (Cleary 1968; Coffman 1961; Pothoff 1966). Consensus upon what to do about the matter is not so readily attained. If a test has been demonstrated to be biased against some particular group, there are three reasonable courses of action to choose from:

(a) the test may no longer be used with that group,

(b) the way in which the test is used may be altered so as to no longer produce unfair outcomes (e.g., by modifying scores, by altering predictions, or by changing interpretations of test results), or

(c) the test may be altered or modified so that it is no longer biased.

The remaining possible course of action, not considered reasonable here, is to ignore the problem and continue to use the test indiscriminately.

There is great diversity of opinion on this point among those writing about this topic. Williams and the Association of Black Psychologists have called for a moratorium on all testing of black children in schools (Williams 1970b); they assume that all tests now in use are biased against blacks and that there is no remedy other than building new, unbiased tests. Neither of these assumptions appears particularly well-founded. First, while one can only guess about bias in any particular test until data have been collected, it is hard to believe that there is much bias, if any, in a number of achievement tests. Tests of skill in arithmetic computation are an example. Also, there is specific evidence that some tests are not biased against blacks. An example is the College Entrance Examination Board's *Scholastic Aptitude Test* which has been shown to work fairly as a predictor of freshman grade point average in so many institutions that it is not reasonable to call it a biased test when so used (Temp 1971). Another example is the CTB/McGraw-Hill *Tests of Basic Experiences* (TOBE) which have been shown to be equally responsive to instruction among black and white kindergarten children (TOBE Bulletin of Technical Data 1971). That is, score patterns of blacks and whites were, if anything, more alike after relevant instruction than before; if the test were biased, the reverse would be true. In short, each test should be separately considered: both blanket endorsements and blanket indictments are inappropriate in view of the small and scattered nature of the evidence to date. Second, in at least some instances, immediate research on the question seems a more appropriate action for a school to take than a cessation of testing, because if bias is discovered it may be possible to adjust the scores or to alter their use to avoid bias to the benefit of all concerned. Sometimes remedies for bias can be found and valid scores produced. Since valid scores are useful, this possibility deserves consideration.

A second position on this issue, equally untenable, is that taken by Messick and Anderson (1970) in their generally excellent article. They acknowledge that some tests may be biased but argue against a moratorium, asserting that the consequences of not testing are worse. It appears they are urging the continued use of biased tests with the very groups against which they are alleged to be biased. They say that people will be more biased without the tests. The logic of this argument is

hard to follow. How can one believe that false and misleading information is better than no information? To hold that it is better to go on using biased tests rather than to stop testing invites the sort of attack Williams has made on the motives of those who build and use tests.

The third position, widely held, does not acknowledge that bias in tests is a serious possibility or problem. Instead, problems of misuse (Munday 1970; Sommer 1970) and/or the inadequate backgrounds of black children are held to be at the heart of the problem (Bennett 1970; Clemans 1970; Milgram 1970; Wikoff 1970). Enough evidence of test bias exists to make this position inappropriate even though one is entitled to believe that upon investigation most tests will prove to be largely unbiased.

A fourth position, preferred here, is that each situation and test should be considered separately. Wherever tests with a substantial probability of bias are in use with one or more groups against whom the tests may be biased, then immediate research on the question should be inaugurated. Where bias is found, adjustments should be made or the records should be discarded. If such research cannot be completed before the results have to be used, then the testing should be postponed or discontinued. There are occasions where unfair outcomes of test use are highly probable because of either test bias or misuse; in such a case it is better not to test. Even here, however, research is in order.

In some instances it will be obvious how such research should proceed; in others it will not. Detailed discussions of possible procedures and some of the difficulties that may arise can be found in a variety of sources (Cardall & Coffman 1964; Coffman 1961; Cleary 1968; Cleary & Hilton 1968; Potthoff 1966; Linn & Werts 1971). It would be misleading to suggest that the matter is simple or fully understood; no comprehensive guide of ways to detect bias in tests is available. Nevertheless, the work should begin.

Messick and Anderson (1970) have sketched briefly the kinds of research that are appropriate. In their excellent section, "The Adequacy of Measurement and the Question of Bias," they note the need for separate assessment of the reliability and validity of a test for each group in order to discount the possibility of bias. Psychologists, schools, and test publishers should now all begin to see to it that these validity studies are made and the results considered before a test is used with any large group which might reasonably be described as different from the majority of the original validation samples.

Since the validation of most tests is a difficult and often a long, drawn-out, expensive, undertaking, it is to be hoped that eventually we can determine the kinds of tests for which such additional validity research is necessary. Right now the possibility of bias is too great to reasonably omit these additional studies for any test intended for general use. There is little appropriate research demonstrating whether or not particular tests are biased, and it will be a long time before statements about tests in general may be expected to suffice.

Williams reports that he has built an ability test, the BITCH test (Black Intelligence Test Counterbalanced for Honkies), which is deliberately biased

in favor of blacks. He notes that blacks score higher on the test than whites, and that he is in the process of determining the validity of this test as a predictor of academic achievement (Williams 1970a). This is a straightforward and unambiguous attack on the problem. If such a test has any validity at all as a predictor of achievement (even of reading comprehension measured with the sorts of passages he offers as appropriate for blacks only) he will have made his point very well. In particular, it should be noted that if the test produces a positive relationship between its scores and some criterion such as school achievement only among, say, urban blacks, then the test constitutes a neat demonstration that biased tests can be built at will and thus supports the contention that other tests are biased against blacks. Some people feel certain that the BITCH test will not prove valid (Wikoff 1970; Munday 1970), but we are a long way from being able to reasonably claim that all such efforts will fail. What Williams is accomplishing with this effort, whether successfully or not as far as the BITCH test is concerned, is a vivid demonstration of the need for tests of ability and achievement to be validated separately for each major group with which they are to be used. More work of this sort is needed.

Of course the investigation of bias in a test should not wait until the test is constructed, much less published. Especially in the case of achievement tests the place to begin is with item tryouts. Evidence about item-by-group interactions is more than just the desirable addition that Messick and Anderson find it; rather, it should become a standard feature of test construction practice unless and until it is demonstrated to be superfluous.

It may be worth noting that this is a recommendation for an increased amount of comparative research. Somewhat inconsistently, Williams objects to comparative research on the grounds that it encourages racist thinking. Since he himself emphasizes the uniqueness of the black psyche and black lifestyle it is difficult to find any logic in this position, but he is clearly right in calling for more research which focuses on the unique resources and strengths of various groups, rather than on "explanations" of the "inferiority" of some group. One good candidate for such research is the question of verbal skills. The kind of verbal skill Williams illustrates in his example of the game "The Dozens" has been remarked upon with great frequency as a highly developed common characteristic of black school children; it appears obvious that this skill has a large cognitive component. Yet many people support the view that "... it is lack of verbal learning..." which is responsible for "... the intellectual and academic deficiencies of disadvantaged children..." (Bereiter & Engelmann 1966, p. 42). If this is true, how does one explain that blacks do relatively better on verbal tests than on other kinds of tests, such as spatial relations tests (Lesser, Fifer, & Clark 1965)? It seems that we have only a dim idea of what verbal skill is all about and what its role is in learning such language-oriented skills as reading (Rystrom 1970). Some better understanding of the possibly unique verbal skills of urban blacks should be most enlightening and more useful than efforts to describe their "linguistic deficiencies."

Unfortunately enlightenment does not occur upon demand, but awareness of a problem is usually the first step. CTB/McGraw-Hill does not claim any breakthrough in dealing with test bias nor does it claim any unusual promptness in becoming fully aware of the problem. However, awareness is here and some steps have been

taken. First, a program of research and exploration has been started. For example, a search for biased items among those in the *California Achievement Tests, 1970 Edition,* is being made with the help of a small grant from the United States Office of Education. Standard item selection procedures are being followed using data from eight different regional and ethnic groups, and performance on the resulting best items for each group are being compared and correlated to see if they are measuring the same thing in all these groups. CTB/McGraw-Hill has also spent a good deal of time and effort in exploring ways in which measurement can be performed more equitably and constructively in schools serving minority groups. In addition to the ameliorative steps, such as using practice exercises and providing training in test-taking skills, we are hopeful that the different sort of conceptual approach to testing represented by criterion-referenced tests will reduce the problem. A number of such instruments are now being developed. The advantages of these tests are:

(a) breadth of coverage,

(b) specific connection to instructional objectives, and especially,

(c) the fact that interpretation does not depend on the summation of diverse performances and does not depend on norms.

Perhaps the different sort of approach to ability measurement referred to by Newland (1970) can accomplish something in that area.

Second, the inclusion of enough black students in item tryouts to search for biased items has been made standard practice in test development. At the moment we do not have enough experience with these kinds of data to know what to expect. If we find most items are biased, in all honesty we would have to either build separate instruments for each ethnic group so differentiated, or recommend against the use of the test with all but one group. The first alternative would probably prove too expensive to be practical and the second would certainly create serious problems for many school systems. It can be readily seen that the accountability movement as well as the national assessment idea would be damaged by such an outcome. Hopefully we will instead find that only a few items show bias; in that case the biased items can simply be eliminated and unbiased tests produced. (It should be remembered that this would not in any way guarantee universality of relevance.)

A different sort of standard of content validity particularly appropriate to criterion-referenced tests is their sensitivity to instruction. That is, do performances on the test change after instruction relevant to the objectives measured? It is CTB/McGraw-Hill policy that this sort of question will be explored for more than just the typical groups used heretofore. In addition, separate studies of the predictive and/or construct validity of new instruments will be undertaken with different groups whenever appropriate and feasible.

A third kind of step, inaugurated several years ago in line with general McGraw-Hill policy, is to have all materials exhibit plainly the multi-ethnic character of American society whenever it is relevant. Fourth, it has been our policy for several years to ensure as best we can that our ability measures are treated as measures of academic aptitude, which they are, not as measures of general intelligence, which they are not. Fifth, we have not hesitated to recommend discontinuing the use of our tests where bias and/or misuse seemed likely and we have maintained vigorously that the use of test scores to exclude children from programs and to deny them opportunities is an egregious error (see, for example, Green 1970).

Finally, we are in the process of exploring whether or not useful improvements in test score interpretation can be obtained by taking into account various characteristics of schools such as neighborhood type, racial composition, size and so forth. Certainly more rational decisions about the value of curricula and programs can be made when such information is available. It also seems possible that test scores modified or interpreted in the light of school variables might have more meaning for instruction than is yielded by the current custom of treating both achievement and aptitude as something a child had or produced independent of his environment.

For the most part the various CTB activities and efforts directed at the issue of test bias are of uncertain value simply because knowledge about the extent and nature of bias in tests is limited. We hope to learn more and hope that all publishers will provide help and cooperation to people like Williams who not only have the courage to attack us but to put their ideas on the line via empirical research.

Bennett, G. K. Response to R. L. Williams. *The Counseling Psychologist*, 1970, 2 (2), 88-89.

Bereiter, C. & Engelmann, S. *Teaching Disadvantaged Children in the Preschool*. Englewood Cliffs, New Jersey: Prentice-Hall, 1966.

Brazziel, W. F. A letter from the South. *Harvard Educational Review*, 1969, 39 (2), 348-356.

Brown, W. M. & Russell, R. D. Limitations of admissions testing for the disadvantaged (letter). *The Personnel and Guidance Journal*, 1964, 43 (3), 301-304.

Cardall, C. & Coffman, W. E. A method for comparing the performance of different groups on the items in a test. *College Entrance Examination Board Research and Development Reports*, 1964, RB 9.

Cleary, T. A. Test bias: prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, Summer 1968, 5, 115-124.

Cleary, T. A. & Hilton, T. L. An investigation of item bias. *Educational and Psychological Measurement*, 1968, 28, 61-75.

Clemans, W. V. A note in response to a request by the editor to comment on R. L. Williams' article entitled, "Black pride, academic relevance, and individual achievement." *The Counseling Psychologist*, 1970, 2 (2), 90-91.

Coffman, W. E. Sex differences in responses to items in an aptitude test. *18th Yearbook, National Council of Measurement in Education*, 1961, 117-124.

Danielian, J. Cognitive ethnocentrism (letter). *Contemporary Psychology*, 1969, 14 (11), 617.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., Tyler, R. W. *Intelligence and Cultural Differences*. Chicago: University of Chicago Press, 1951.

Green, D. R. Letter to South Carolina Community Relations Program, American Friends Service Committee. *Your Schools*, 1970, 1 (9).

Lesser, G. S., Fifer, G., & Clark, D. H. Mental abilities of children from different social-class and cultural groups. *Monographs of the Society for Research in Child Development*, 1965, 30 (4, Whole No. 102).

Linn, R. L. & Werts, C. E. Considerations for studies of test bias. *Journal of Educational Measurement*, Spring 1971, 8 (1), 1-4.

Marks, E. & Noll, G. A. Procedures and criteria for evaluating reading and listening comprehension tests. *Educational and Psychological Measurement*, Summer 1967, 27 (2), 335-348.

Milgram, N. A. Danger: chauvinism, scapegoatism, and euphemism. *Clinical Child Psychology Neswletter*, 1970, 9 (3).

Messick, S. & Anderson, S. Educational testing, individual development, and social responsibility. *The Counseling Psychologist*, 1970, 2 (2), 80-88.

Munday, L. A. Measurement for equal opportunity. *The Counseling Psychologist*, 1970, 2 (2), 93-97.

Newland, T. E. Testing minority group children. *Clinical Child Psychology Newsletter*, 1970, 9 (3), 5.

Pintner, R. *Intelligence Testing*. New York: Henry Holt & Co., 1923.

Potthoff, R. F. Statistical aspects of the problem of biases in psychological tests. *University of North Carolina Institute of Statistics Mimeo Series*, 1966, No. 479.

Rystrom, R. Dialectical training in reading: a further look. *Reading Research Quarterly*, 1970, 5, 581-599.

Sommer, J. Response to R. L. Williams. *The Counseling Psychologist*, 1970, 2 (2), 92.

Stanley, J. C. Predicting College success of the educationally disadvantaged. *Science*, 1971, 171 (3972), 640-647.

Temp, G. Test bias: validity of the S.A.T. for blacks and whites in thirteen integrated institutions. *College Entrance Examination Board Research and Development Reports*, 1971, RB. 712.

Tests of Basic Experiences Bulletin of Technical Data. Monterey, California: CTB/McGraw-Hill, 1971.

Wasserman, M. Planting pansies on the roof. *The Urban Review*, 1969, 3 (3), 30-35.

Wikoff, R. L. Danger: attacks on tests unfair. *Clinical Child Psychology Newsletter*, 1970, 9 (3), 3-4.

Williams, R. L. Danger: testing and dehumanizing black children. *Clinical Child Psychology Newsletter*, Spring 1970, 9 (1), 5-6. (a)

Williams, R. L. Black pride, academic relevance and individual achievement. *The Counseling Psychologist*, 1970, 2 (1), 18-22. (b)

_____. From dehumanization to black intellectual genocide: a rejoinder. *Clinical Child Psychology Newsletter*, 1970, 9 (3). (c)

Wrightstone, J. W. As Dr. Wrightstone sees it (letter). *The Urban Review*, 1969, 4 (1), 45-47.